*Nucleic Acids Research, 2004, Vol. 32, Web Server issue* ***W267–W272***
*DOI: 10.1093/nar/gkh473*

# KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains

**Kiyoko F. Aoki*, Atsuko Yamaguchi, Nobuhisa Ueda, Tatsuya Akutsu, Hiroshi Mamitsuka, Susumu Goto and Minoru Kanehisa**

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Received February 13, 2004; Revised April 20, 2004; Accepted May 3, 2004

## ABSTRACT

**KCaM (KEGG Carbohydrate Matcher) is a tool for the analysis of carbohydrate sugar chains, or glycans. It consists of a web-based graphical user interface that allows users to enter glycans easily with the mouse. The glycan structure is then transformed into our KCF (KEGG Chemical Function) file format and sent to our program which implements an efficient tree-structure alignment algorithm, similar to sequence alignment algorithms but for branched tree structures. Users can also retrieve glycan tree structures in KCF format from their local computers for visualization over the web. The tree-matching algorithm provides several options for performing different types of tree-matching procedures on glycans. These options consist of whether to incorporate gaps in a match, whether to take the linkage information into consideration and local versus global alignment. The results of this program are returned as a list of glycan structures in order of similarity based on these options. The actual alignment can be viewed graphically, and the annotation information can also be viewed easily since all this information is linked with KEGG's comprehensive suite of genomic data. Analogously to BLAST, users are thus able to compare glycan structures of interest with glycans from different glycan databases using a variety of tree-alignment options. KCaM is currently available at http://glycan.genome.ad.jp.**

## INTRODUCTION

Oligosaccharides, or glycans, have only recently started gaining attention from the bioinformatics community. This recent advance may be attributed to better technologies and a better understanding of glycan structures. Unlike the linear structures of DNA and proteins, glycans are tree structures, where one monosaccharide unit may be connected to one or more other monosaccharide units. The linkages between these units also have variables, such as the anomer (α or β) and the hydroxyl group numbers to which they are attached on the monosaccharides (1).

When analyzing new sequences, biologists often turn to programs such as BLAST (2) to perform sequence alignment with other known sequences, usually stored in a database. For tree structures, however, the same programs cannot be applied; straightforward sequence alignment algorithms would not suffice when analyzing glycan structures. Also, although other carbohydrate databases are available on the web, their query interfaces are not exactly intuitive in that the glycan tree structures often have to be entered textually (3–5). Thus, we were prompted to develop KCaM (KEGG Carbohydrate Matcher), which is a user-friendly interface to an efficient program that performs glycan tree-structure alignments on different databases via the web. By making available a web-based server that (i) provides a Java-based interface for biologists to easily specify glycan structures with their mouse, (ii) sends this structure in a standardized format to the algorithms of KCaM and (iii) visualizes the resulting alignments, KCaM makes it possible for users to easily align and further analyze any glycan structure over the web.

## MATERIALS AND METHODS

### The KCaM algorithms

KCaM consists of two main variations, an approximate matching algorithm and an exact matching algorithm. The approximate matching algorithm aligns monosaccharides while allowing gaps in the alignment, and the exact matching algorithm aligns linkages while disallowing any gaps, thus resulting in a stricter criterion for alignments. Both variations provide local and global options, just as in sequence alignment

*To whom correspondence should be addressed. Tel: +81 774 31 4902; Fax: +81 774 31 4904; Email: kiyoko@kuicr.kyoto-u.ac.jp algorithms. The approximate matching algorithm does not penalize gaps for unaligned regions in the local option (as opposed to the global option) such that only conserved regions can be found. The global option in the exact matching algorithm is a recursive version of the local option, which runs the exact matching algorithm once. The following rules of thumb may be used when deciding on which algorithm to use.

*Approximate matching.* The approximate matching algorithm is useful for the case when a core structure is known, and we are looking for a specific pattern near the leaves. The local or global version can then be chosen based on the size of the tree.

*Exact matching.* When looking for a single connected subtree, the exact matching algorithm that allows no gaps may be preferred. In this case, the local version can be selected to find one specific subtree, or the global option can be used to find as many matches as possible. For the exact version, the degree of specificity can be specified by selecting either to match just monosaccharide names ('Sugar Only') or both names and linkage information ('Sugar & Bond').

## KCaM calculations

The alignment algorithm of KCaM follows that of the Smith–Waterman algorithm (6,7), except that it has been modified for tree structures. Given that the average number of monosaccharides in a glycan structure is <10, this procedure is both efficient and accurate. The details of the dynamic programming procedures for both the local and global versions are explained in (8).

Figure 1 illustrates the workflow of the KCaM web server system. Once the query structure is entered, the database from which to query is selected, and the KCaM options are specified, the query is sent to the server. For each entry in the selected database, KCaM takes into account the multiple monosaccharides and linkages from each monosaccharide, so in step (i) it organizes the trees such that the monosaccharides can be traversed in order, from root to leaves. In step (ii), it calculates the similarity between the monosaccharides (see the next section) at the farthest branches of each tree and traverses towards the root, combining the similarity calculations as it goes. Once at the root, depending on whether a local or global calculation is being made, in step (iii) either the value at the

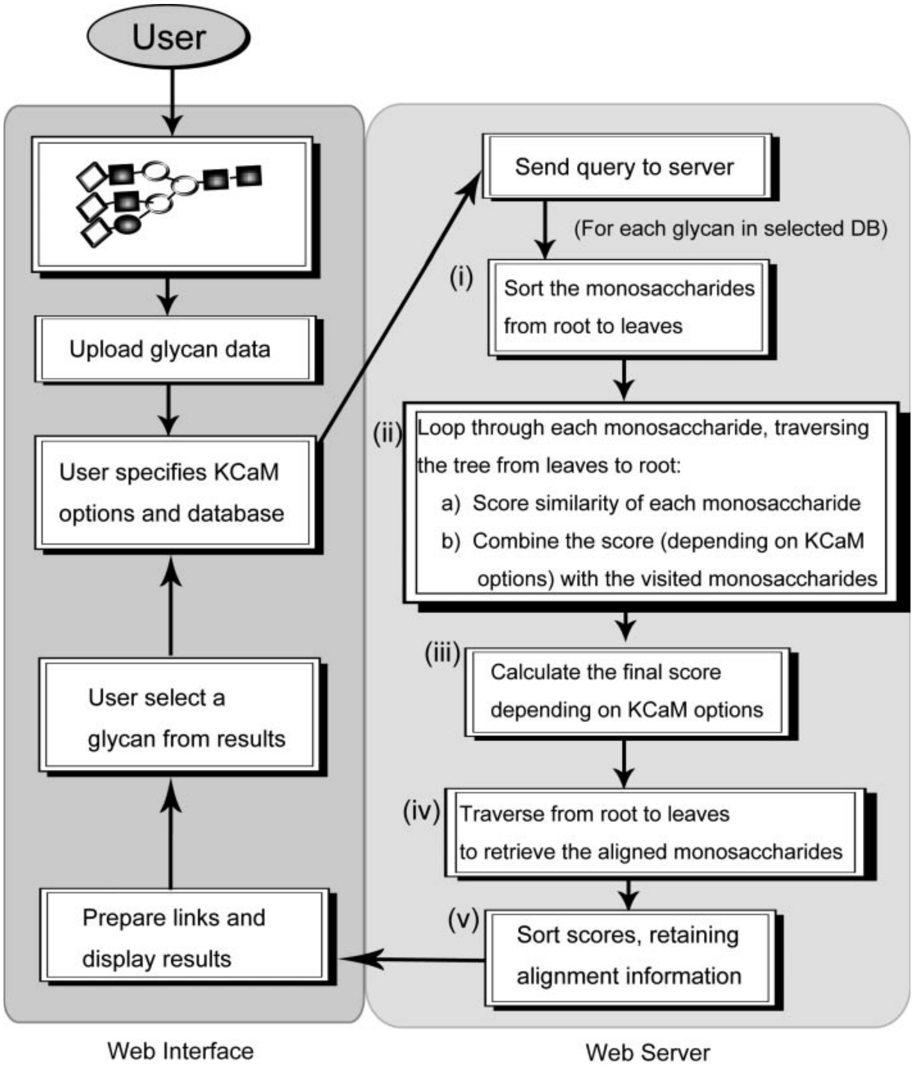
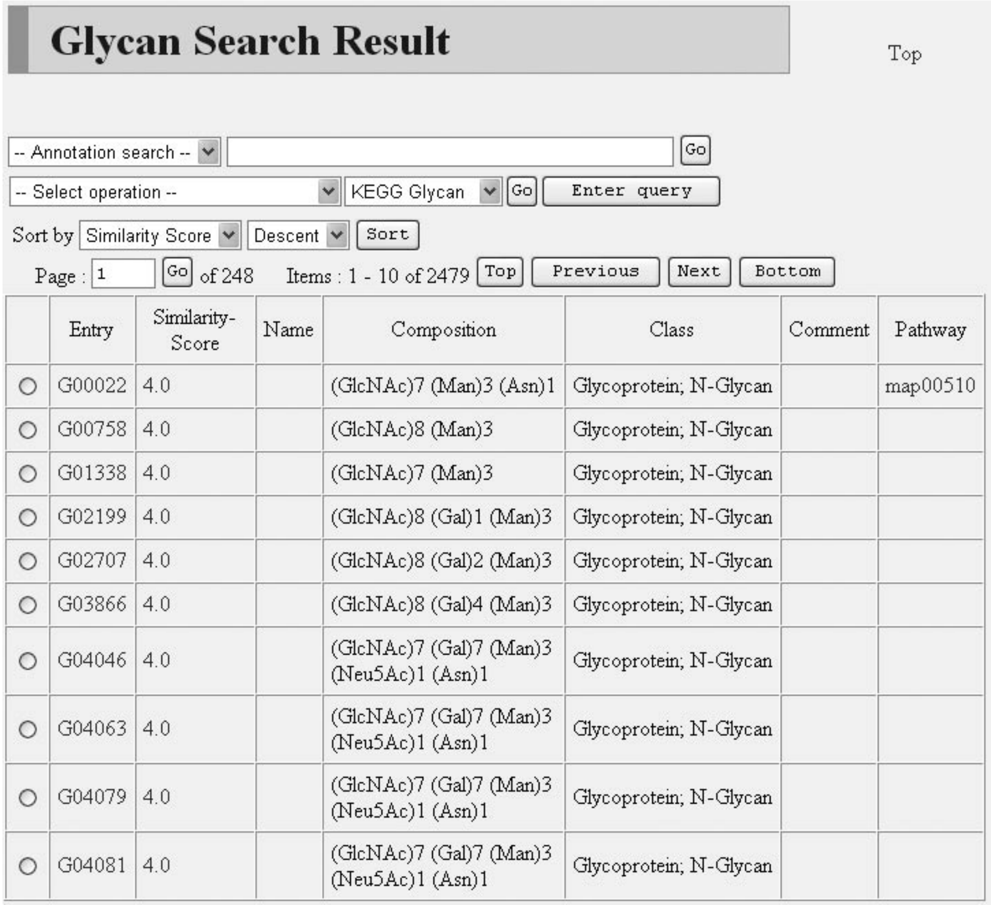

User
Upload glycan data
User specifies KCaM options and database
User select a glycan from results
Prepare links and display results
Send query to server
(For each glycan in selected DB)
(i)
Sort the monosaccharides from root to leaves
(ii)
Loop through each monosaccharide, traversing the tree from leaves to root:
a) Score similarity of each monosaccharide
b) Combine the score (depending on KCaM options) with the visited monosaccharides
(iii)
Calculate the final score depending on KCaM options
(iv)
Traverse from root to leaves to retrieve the aligned monosaccharides
(v)
Sort scores, retaining alignment information
Web Interface
Web Server

**Figure 1.** An illustration of the KCaM server workflow.



Sort by Similarity Score Descent

Page : 1 of 248 Items : 1 - 10 of 2479

| | Entry | Similarity-Score | Name | Composition | Class | Comment | Pathway |
|---|---|---|---|---|---|---|---|
| ○ | G00022 | 4.0 | | (GlcNAc)7 (Man)3 (Asn)1 | Glycoprotein; N-Glycan | | map00510 |
| ○ | G00758 | 4.0 | | (GlcNAc)8 (Man)3 | Glycoprotein; N-Glycan | | |
| ○ | G01338 | 4.0 | | (GlcNAc)7 (Man)3 | Glycoprotein; N-Glycan | | |
| ○ | G02199 | 4.0 | | (GlcNAc)8 (Gal)1 (Man)3 | Glycoprotein; N-Glycan | | |
| ○ | G02707 | 4.0 | | (GlcNAc)8 (Gal)2 (Man)3 | Glycoprotein; N-Glycan | | |
| ○ | G03866 | 4.0 | | (GlcNAc)8 (Gal)4 (Man)3 | Glycoprotein; N-Glycan | | |
| ○ | G04046 | 4.0 | | (GlcNAc)7 (Gal)7 (Man)3 (Neu5Ac)1 (Asn)1 | Glycoprotein; N-Glycan | | |
| ○ | G04063 | 4.0 | | (GlcNAc)7 (Gal)7 (Man)3 (Neu5Ac)1 (Asn)1 | Glycoprotein; N-Glycan | | |
| ○ | G04079 | 4.0 | | (GlcNAc)7 (Gal)7 (Man)3 (Neu5Ac)1 (Asn)1 | Glycoprotein; N-Glycan | | |
| ○ | G04081 | 4.0 | | (GlcNAc)7 (Gal)7 (Man)3 (Neu5Ac)1 (Asn)1 | Glycoprotein; N-Glycan | | |

**Figure 2.** Resulting page of similar (best aligned) structures to the input structure. The number of resulting structures is listed, and each entry is provided with links to display its annotation information as well as alignment details.

root or the maximum value across the tree is returned as the final score. To retrieve the actual monosaccharides contributing to the final score, in step (iv) they are traversed back down the tree and tracked. The final set of monosaccharides and linkages is then returned along with the final score. Step (v) sorts the scores for all structures aligned with the query and returns all relevant alignment information for display.

## KCaM scores

The scoring schemes currently used in KCaM are different for the approximate and exact matching versions because of the calculation of gaps. The approximate matching algorithm measures the similarity between two monosaccharides based on the following formula:

$$\begin{aligned} w(u,v) = \max[0, \alpha\delta[\text{label}(u), \text{label}(v)) \\ - \beta(1 - \delta(\text{ulabel}(p(u),u], \text{ulabel}(p(v),v))) \\ -\beta(1 - \delta(\text{dlabel}(p(u),u), \text{dlabel}(p(v),v)))], \end{aligned}$$

where $\delta(x, y) = 1$ iff $x = y$, label($u$) denotes the name of the monosaccharide and ulabel($p(u)$, $u$) [respectively dlabel ($p(u)$,$u$)] indicates the hydroxyl group of monosaccharide $p(u)$ (respectively monosaccharide $u$) of the linkage ($p(u)$,$u$).

The exact matching algorithm returns as its score the number of linkages that were aligned between the two trees. In other words, exactly matching linkages, including both monosaccharide names and hydroxyl groups, receive scores of 1.0; otherwise the score is 0.0.

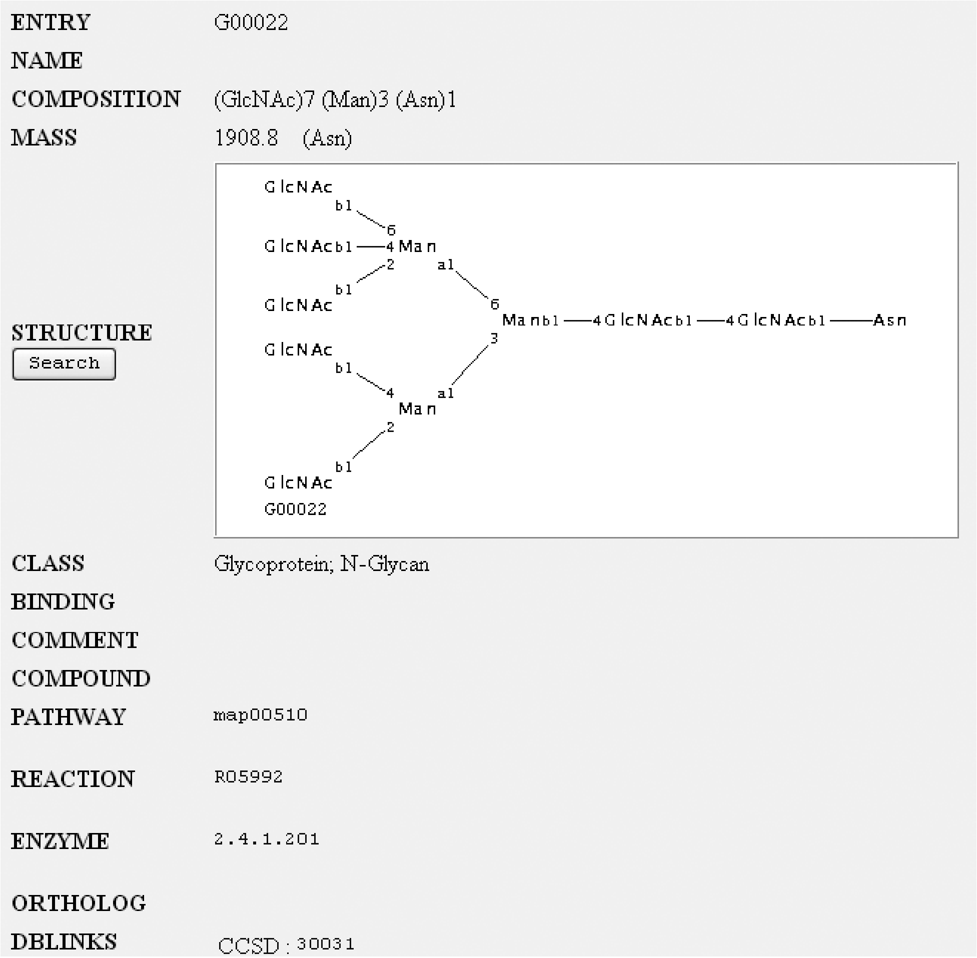
## The KCaM main page

KCaM provides a database option available for those who wish to perform an alignment with the structures in either the latest KEGG Glycan database (9) or the original CarbBank/CCSD (10) database. The KCaM options are listed under 'Search Type' on the main web page.

The user can then enter the actual glycan structure they wish to use as a query structure by clicking on the 'Click here . . . ' link. This will bring up a new window with the glycan structure editor, described in the next section. The main web page also includes links to a Frequently Asked Questions (FAQ) section as well as a tutorial for new users to quickly familiarize themselves with KCaM.

## The glycan structure editor

The glycan structure editor (S. Goto *et al*., manuscript in preparation) is a user-friendly interface for entering any glycan structure over the web. It has been modeled after CambridgeSoft's ChemDraw$^{TM}$ and MDL Information



| | |
|---|---|
| ENTRY | G00022 |
| NAME | |
| COMPOSITION | (GlcNAc)7 (Man)3 (Asn)1 |
| MASS | 1908.8 (Asn) |
| STRUCTURE<br>Search | |
| CLASS | Glycoprotein; N-Glycan |
| BINDING | |
| COMMENT | |
| COMPOUND | |
| PATHWAY | map00510 |
| REACTION | R05992 |
| ENZYME | 2.4.1.201 |
| ORTHOLOG | |
| DBLINKS | CCSD : 30031 |

**Figure 3.** Annotation information for entry ID G00022. All available pathway, reaction, enzyme and other information related to the entry is available via links.

Systems Inc.'s ISIS$^{TM}$/Draw to minimize the learning curve for its usage. Users may either upload text files in KCF (KEGG Chemical Function) format from their local computers or input a glycan structure directly into the applet. The glycan specified thus is used as the query structure in the alignments.

The glycan structure editor has been tested on several browsers and platforms as listed on the FAQ page, and is fully functional with Internet Explorer 6 on Windows 2000/XP, Netscape 7.02 on Mac OS X (10.3.2) and Vine on Linux, among others. There are currently known issues with Internet Explorer 5.2 and Opera 6.03 on MacOS X (10.3.2), Internet Explorer 5 on Mac OS 9.2 and Netscape 7.02 on Windows XP.

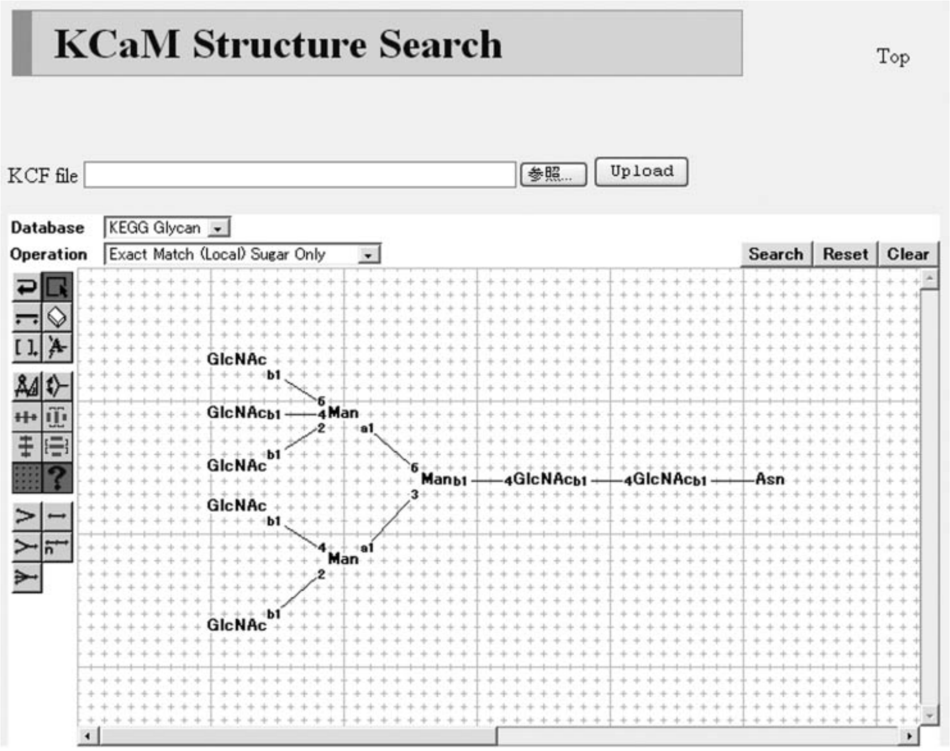
Once an alignment query structure has been entered, KCaM can be invoked with the selected parameters and the given query structure, resulting in a page as in Figure 2.

## RESULTS

The results of KCaM provide a plethora of information useful for analyzing the glycan structures related to the entered structure. An indication of the number of results from the selected database that matched the structure most closely is provided, and the results are then listed below in order of similarity, in groups of 10 entries per page.

For each resulting structure, its entry ID, similarity (or alignment) score, name, composition, class, comments and pathway information are provided. The entry ID is linked to the full annotation information associated with the given entry, a sample of which is illustrated in Figure 3. Any information available related to the given entry is provided, along with direct links to the sources of information. In the example in Figure 3 for entry ID G00022, we see that this structure is an N-linked glycan involved in the N-Glycan biosynthesis pathway. A link is provided that would lead to the page for the actual KEGG pathway. We also see that it is involved in a reaction whose ID is R05992 from the KEGG Reaction database, and it is related to the enzyme with EC number 2.4.1.201. Each of these links will launch a new page with the respective database entries. Other information (not shown), such as PubMed IDs, is also available and linked. Finally, we see that this entry was originally documented from CCSD IDs 6926 and 19072. The user can click on the 'CCSD' link to see both of these entries together or on each individual link to see them separately.

This page also allows the user to click on the glycan image (or the Search button) to use this entry as the query for another alignment. When the image is clicked, a new



KCaM Structure Search

KCF file

Database: KEGG Glycan

Operation: Exact Match (Local) Sugar Only

**Figure 4.** A new search using the selected entry (or a modification of it) can be made.

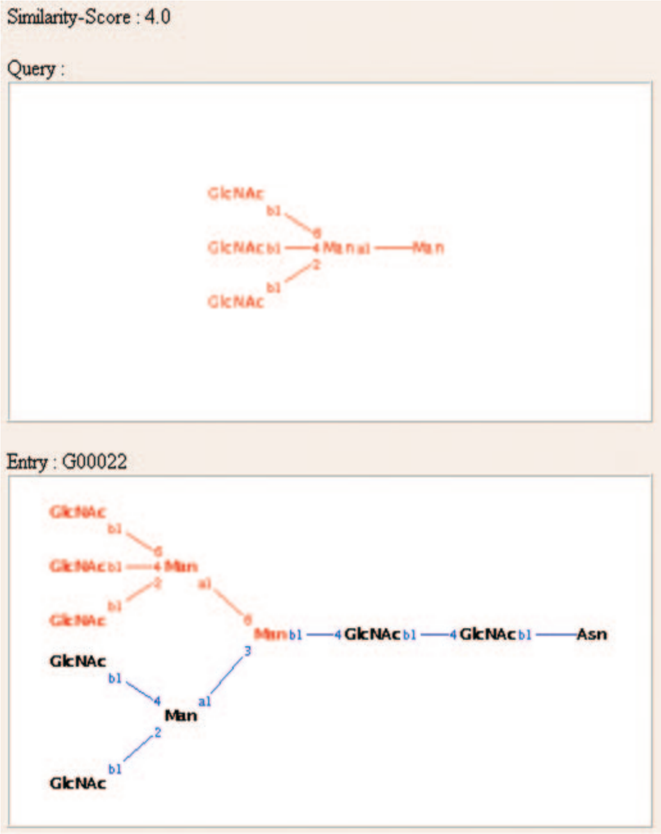
window as in Figure 4 will be displayed. The user is then free to use this structure or modify it to perform a new glycan query.

Returning to the original search results page, the pathways in which the resulting structures are involved are linked, and we can also click on the Similarity Score value to see how the structures were aligned, as in Figure 5, which displays the actual alignment result between the query structure and entry G00022. The red portions indicate the aligned portions of the structures that contributed to the score.

Finally, we note that the results from this search can also be used as a subdatabase for further refined searches. The drop-down box at the top can be used as in the first main page to select the textual criterion with which to search repeatedly, or the drop-down box in the second row can be used to select one of the resulting structures (selected by clicking on the corresponding entry's radio button to its left) as a new query structure. Finally, the resulting list may be sorted by selecting various options, available in the third row from the top.

## DISCUSSION

Similarity of glycan structures is suspected to be related to glycan function (11,12). Thus, it is expected that the alignments provided by KCaM will be useful for glycobiologists, just as sequence alignments are used for numerous applications in DNA and protein research. The web-based interface for KCaM allows glycobiologists to easily align their glycan structures to other known structures available in the KEGG



Similarity-Score : 4.0

Query :

Entry : G00022

**Figure 5.** Graphical display of the aligned structures, with the alignment indicated in red.

Glycan and CarbBank/CCSD databases. Owing to the abundance of data available via KEGG's various database resources, biologists are provided with a comprehensive set of information for use with their analysis. KCaM is also practical in that its various options provide for a variety of needs, and its graphical user interface is user-friendly in terms of both glycan structure entry and visualization.

Work is continuing on improving the scoring schemes used in KCaM. A detailed analysis of glycan alignments is currently under way and will be implemented in the near future to take into account the frequency of occurrence of certain types of linkages in glycans and their corresponding alignments. By analyzing the frequency of occurrence of the linkages in the actual alignments, a score matrix can be developed to provide more meaningful scores. This would also allow KCaM to provide confidence scores such as *E*-values, similarly to BLAST, to provide an indication of the significance of the alignments. Future work will also consist of incorporating chemical compound structures into the glycan structures for more complex searches. Such functionality would provide users with a method to integrate richer information into their queries, thus retrieving more meaningful information in their results.

We note here that KCaM is still in its early stages of development at this time, as is glycome informatics. Work is in progress to improve the algorithms, and the KEGG Glycan data itself is continually being curated to provide glycobiologists with the most useful information. KCaM is a potentially very useful tool for this growing field of glycome informatics, and as improvements are made, it is expected to gain more usage in the near future.

## ACKNOWLEDGEMENTS

We would like to acknowledge Masami Hamajima, Masayuki Kawasima and Shigekazu Masumoto for their support in providing the Java-based interface for KCaM. This work is supported in part by Grant-in-Aid for Scientific Research on Priority Areas (c) ''Genome Information Science'' from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## REFERENCES

1. Varki,A., Cummings,J., Esko,J., Freezie,H., Hart,G. and Math,J. (eds) (1999) *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Bunsmann,P., Bohne,A., Loss,A., Schwarzer,E., Lang,E. and von der Lieth,C.W. (2002) SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res.*, **30**, 405–408.
4. van Kuik,J. and Vliegenthard,J.F. (1992) Databases of complex carbohydrates. *Trends Biotechnol.*, **10**, 182–185.
5. van Kuik,J., Hard,K. and Vliegenthart,J.F. (1992) A $^1$H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr. Res.*, **235**, 53–68.
6. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
7. Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.
8. Aoki,K.F., Yamaguchi,A., Okuno,Y., Akutsu,T., Ueda,N., Kanehisa,M. and Mamitsuka,H. (2003) Efficient tree-matching methods for accurate carbohydrate database queries. *Genome Informatics*, **14**, 134–143.
9. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
10. Doubet,S., Bock,K., Smith,D., Darvill,A. and Albersheim,P. (1989) The complex carbohydrate structure database. *Trends Biochem. Sci.*, **14**, 475–477.
11. Bertozzi,C.R. and Kiessling,L.L. (2001) Carbohydrates and glycobiology review: chemical glycobiology. *Science*, **291**, 2357–2364.
12. Drickamer,K. (1988) Two distinct classes of carbohydrate-recognition domains in animal lectins. *J. Biol. Chem.*, **263**, 9557–9560.